Speech recognition for program control and data entry in a production environment

Susan E. Hauser*, Tehseen F. Sabir, George R. Thoma

Lister Hill National Center for Biomedical Communications
National Library of Medicine, Bethesda, MD 20894

ABSTRACT

The Lister Hill National Center for Biomedical Communications, an R&D division of the National Library of Medicine, has developed a PC-based system for semi-automated entry of journal citation data into MEDLINE®. The system, called MARS for Medical Article Records System, includes many automated features but requires a few manual tasks such as scanning and the entry of certain data that are not located on the scanned page. Now that considerable computing power and speed are routinely available on desktop PCs, we think it may be possible to include speech recognition as an optional user interface to reduce operator burden and to improve speed and quality for document scanning and data entry. We undertook a study to determine if speech recognition was sufficiently accurate, reliable and immune to noise to warrant integration with MARS workstations.

The study focussed on the suitability of both continuous and discrete speech recognition for computer program control and for data entry in a production environment. Following a carefully structured format, 20 participants tested continuous speech recognition and 20 participants tested discrete speech recognition in both a quiet and a noisy environment. Performance measures were accuracy and speed. Both continuous and discrete recognition were very accurate, fast and immune to noise when used for program control. For data entry, though discrete speech recognition was about 90% accurate, it was very slow. Continuous speech recognition was faster for data entry, but was only about 76% accurate. As a result of the study, speech recognition has been integrated into the MARS scan workstation for program control.

Keywords: Speech recognition, software evaluation, process control, National Library of Medicine

1. INTRODUCTION

1.1. Background

The National Library of Medicine (NLM) is automating the production of bibliographic records for its premier database of bibliographic records, MEDLINE. As a first step, the Communications Engineering Branch (CEB) of the Lister Hill National Center for Biomedical Communications has developed a system called MARS, for Medical Article Record System[1]. The first version, MARS-1, involves scanning and converting by optical character recognition (OCR) the abstracts that appear in journal articles, while keyboarding the remaining fields (e.g., article title, authors, affiliations, etc). We are designing and developing the second generation system MARS-2 to automate the entry of these other fields as well. However, human operators will still be required for certain operations such as scanning and the entry of special data fields.

The exponential increase of computing power available in personal computers permits desktop applications that were once relegated to research institutions and large companies. One such application is speech recognition. The past few years has seen the introduction of several commercial speech recognition packages for PC and Macintosh platforms[2,3,4,5]. These products have proven useful in specialized "hands-

* Contact: http://archive.nlm.nih.gov/staff/hauser.php

busy, eyes-busy" environments [6], or to computer users who have physical difficulty with conventional input devices [7]. Other attempts to apply speech recognition directly to existing computer interactive applications have had mixed results [8,9]. Success with currently available speech recognition packages seems more likely if the speech recognition application is carefully chosen [10,11], and then integrated into the existing computing environment. An early study by Karl et al [6] found that speech recognition enhanced user interfaces when it served as an additional input channel, supplementing the keyboard and mouse, for short, highly interactive transactions.

1.2. Objective

Speech recognition has improved dramatically over the past few years, but is still relatively unproven for PC-based applications such as MARS. Few well conducted performance evaluations of commercially available technical products are available for reference. The reviews of PC-based speech recognition products that appear in the popular press tend to be positive for specific applications but with reservations about general-purpose use. To avoid introducing problems by implementing a new technology into a production environment, it was necessary to conduct our own application-specific evaluation of speech recognition technology.

Two off-the-shelf speech recognition products were selected for evaluation. Both products are reasonably priced and run on a PC platform under either Windows or NT. Both are widely used within the community of computer users who are motivated by pain or incapacity to continuously train and maintain the software[12]. Our target users would not be so motivated. Being production oriented, they would have little tolerance for training, vocabulary maintenance or even an occasional failure. Therefore speech recognition in our application would have to work reliably with only a short period of training. However, because the size of the vocabulary needed to accomplish our computer control and data entry tasks is small, we expect these products to prove satisfactory without a large investment of training time on the part of the users. And because we have very application-specific requirements and can tailor speech recognition to our environment, we anticipate an appropriate niche for the current capabilities of Speech Recognition technology.

1.3. Significance

Although the next version of MARS will include many automated features, it will still require a few manual tasks such as scanning and the entry of certain data types. The first step in the MARS-2 process is to scan the first page of every article in the journal being processed. To be cost effective, the scanning operation must be fast and produce good quality images on the first scan. Unlike the MARS-1, where only the article title and abstract were converted by OCR, in MARS-2, the entire area of each scanned page will be converted. To obtain acceptable image quality for the whole page, often both hands are required to apply even pressure to the page being scanned. By controlling the scan operation by voice, the operator would have both hands free to manipulate the journal, thus ensuring a better quality image. This may also improve throughput because the operator need not reach for the mouse or keyboard, and can concentrate on the correct positioning of the journal on the scanner platen.

Only the first page of each article is scanned for OCR conversion, since the principal fields generally appear on the first page. However, certain MEDLINE citation data, such as NIH Grant numbers and Databank Accession numbers are not always located on the first page. When present, these relatively long alphanumeric strings will be entered manually. By entering these numbers via speech recognition, the operator could have both hands free to manipulate the journal, and possibly be able to enter the numbers with greater accuracy or speed than by typing them.

In summary, it is anticipated that integrating speech recognition into selected MARS-2 workstations as an optional user interface may reduce operator burden and improve speed and quality of document scanning and data entry. If proven successful in these applications, it can be introduced in other workstations to support verification.

2. METHODS

Two speech recognition technologies were evaluated. Discrete Speech Recognition technology was evaluated using Dragon Dictate 3.0 [2] and Continuous Speech Recognition technology was evaluated using Dragon Naturally Speaking 2.02 [2]. Discrete Speech Recognition requires the speaker to pause slightly between each word, while Continuous Speech Recognition permits the speaker to speak "naturally", without such pauses. Both of these products are speaker-dependent, i.e., each speaker must train the system for his or her voice. Since our intended applications of Speech Recognition include very little actual "continuous" speech, it may have been sufficient to only test Discrete Speech Recognition Technology. We thought, however, the differences in the user interface and underlying technologies of the two products may confer an advantage to Continuous Speech Recognition technology which should be tested in our environment.

The test platform consisted of a Pentium Pro 166 MHz CPU, 64 MB of RAM, a Creative Labs SoundBlaster AWE64 Gold and a VXI headset. To the extent possible, the study attempted to evaluate the two speech recognition technologies rather than Dragon Systems products. The tests were designed so that the participants would not need to learn the details of using either product.

Evaluation formats were developed to model the two likely uses for speech recognition in the actual system: *Program Control* and *Data Entry*. The Program Control format included commands used in the scanning operation. The Data Entry format included strings of words, letters and numbers found in NIH Grant Numbers and Databank Accession numbers. Since the objective was to test speech recognition, not Dragon Dictate or Dragon Naturally Speaking, the tests were designed to isolate our participants from the user interface details that were particular to Dragon Systems' products. We provided exact scripts for each participant to read, including the use of the military alphabet (to enter alphanumeric strings), and we instructed the participants to not correct any errors made by the recognition system. To the extent possible we also designed the tests to be identical for each technology. Although our participants did need to train the technology they were testing, they were isolated from other differences in vocabulary management issues.

Twenty participants, ten women and ten men, were recruited to evaluate each of the two technologies. The participants represented a broad range of ages, accents and voice types. Before testing, the participants trained the appropriate software to recognize their voices. In addition to the standard training recommended by the manufacturer, the participants trained the system for additional special command words and short phrases that would be used in the subsequent tests. The Discrete Speech Recognition program required about 20 minutes of training; the Continuous Speech Recognition program required about 50 minutes of training. The training session not only prepared the system for the participants' voices, it also let the participants become familiar with the concept of talking to a computer. The participants were permitted to break for several hours or days between the training session and the testing session. Although both the Continuous and Discrete Speech Recognition programs permit, and even encourage, real-time correction while dictating, the test participants did no correcting, as this would have required additional effort on the part of every participant. Every effort was made to ensure that it was the system that was being tested, not the participant.

Each participant tested either the Discrete Speech Recognition program or the Continuous Speech Recognition program in four scenarios: Program Control in a quiet environment, Program Control in a noisy environment, Data Entry in a quiet environment and Data Entry in a noisy environment. To reduce the effect of practice and familiarity, the order of testing in a quiet or noisy environment was randomly assigned. The "quiet" environment was a typical office cubicle, with a measured background noise of 38.4 dBA. A "noisy" production environment was simulated in the same cubicle by playing a recording taken in the MARS-1 production area, in the noisy area near the main printer and several keyboard operators. The noise level measured at the participant's microphone during playback was 56.4 dBA. This level exceeded the ambient noise level in the production area by about 7 dBA [13].

Figure 1 is an example of what the participant sees when testing Program Control. At the beginning of the test, the right side of the screen is blank. The participant reads a total of 31 commands that appear in the **left** column, each on a separate line. To provide feedback during testing [6], when the software recognizes a command, it invokes a macro to insert the text of the recognized command on the **right** side of the screen and move the cursor down. Once the text appears on the right side of the screen, the participant can proceed to the next command. In the situation shown here, the next thing that the participant would say is “scan and insert”. Even when using Discrete Speech Recognition, the participant can say "scan and insert" rather than "scan (pause) and (pause) insert" because part of the training was to teach the system to recognize the whole phrase "scan and insert" as one utterance. The first command shown is a special command, TS command, whose associated macro inserts a timestamp into the output file as shown in Figure 1. This command is used at the beginning and ending of portions of the tests in order to get a rough measure of speed. The participant was instructed to say each command only once and make no attempt to correct mistakes made by the recognition package. For Program Control evaluation, the test scenario was the same for both technologies.

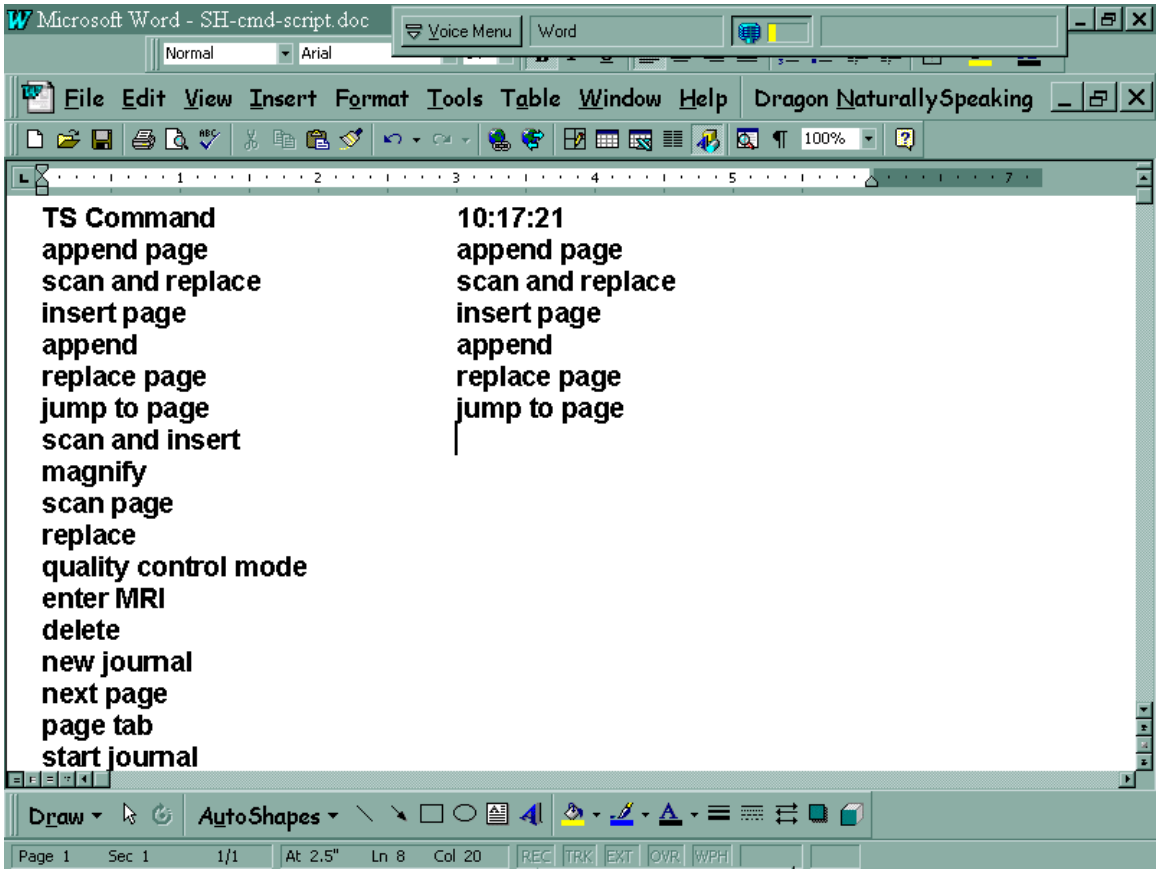

Figure 1. What the Program Control tester sees.

Figure 2 is an example of what the participant sees when testing the entry of Databank Accession numbers. The participant reads 20 lines of words, letters and numbers from the bold lines on the **left** side of the screen. For example: "swissprot papa enter number 80935". Each line is a Databank Accession number from actual MEDLINE records. To facilitate entry of the Databank names, the words 'swissprot." "genbank" and "PDB" (pronounced peedeebee) were added to the active vocabulary. We also used the military alphabet to maximize recognition accuracy, but we provided each of those words for the participants to read; they were not expected to memorize the alpha-bravo terms. The recognized words are immediately inserted on the **right** side of the screen. The line below each bold line lets the speaker see what the result is supposed to look like, even though the participant does no correction. The Databank Accession numbers are displayed in groups of ten, with an opportunity to rest between each group. The format shown in Figure 2 is testing Discrete Speech Recognition. The phrase "enter number" appears several times. This allows the speaker to say a series of numbers without pausing between each number, as would be the default requirement in Discrete Speech Recognition. The test scenario for Databank Accession numbers was **almost** identical for the two technologies. The only difference was that for Continuous Speech Recognition, the participant would **not** say the phrase "enter number" before saying a series of numbers.

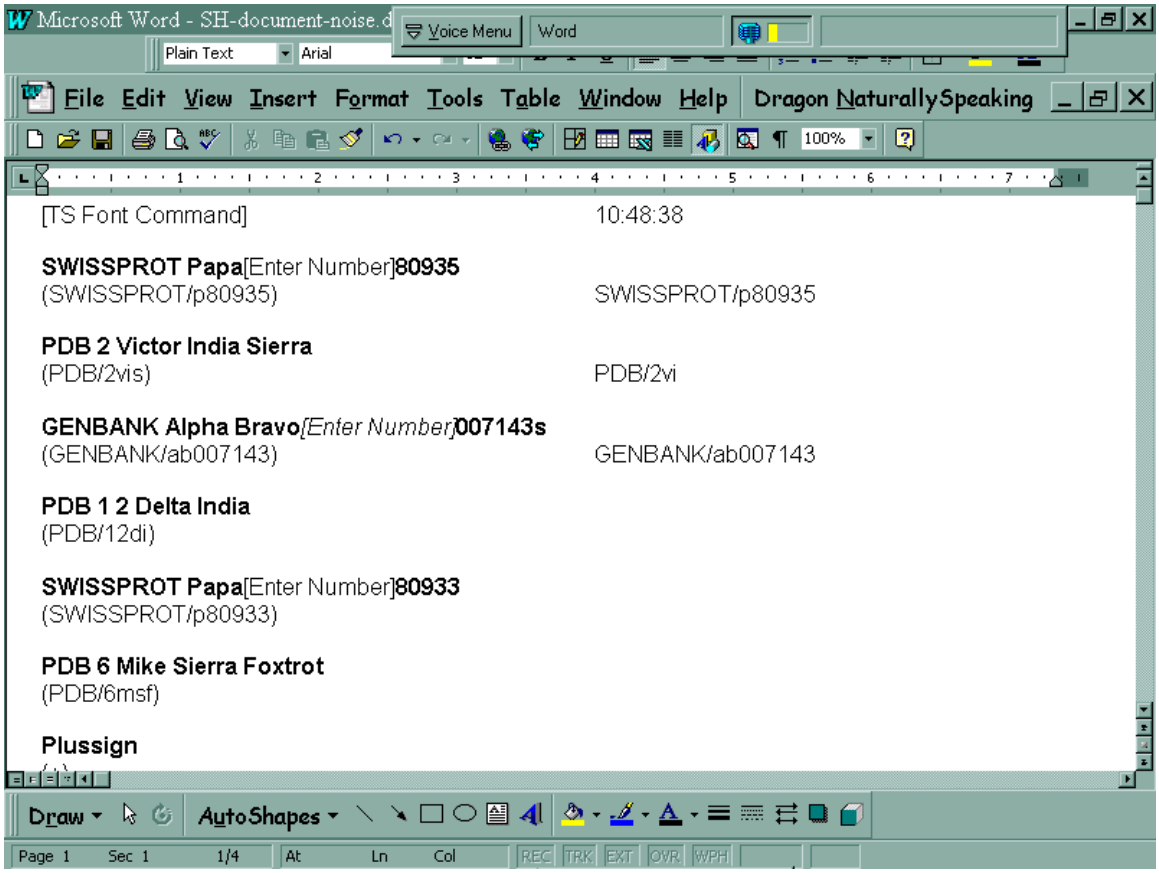

Figure 2. What the Data Entry tester sees when testing Databank Accession Numbers.

Figure 3 is an example of what the participant sees when testing the entry of NIH Grant numbers. It is similar to the format for testing Databank Accession numbers. The participant reads a total of 20 NIH Grant numbers, with an opportunity to rest after each group of ten. To minimize the testers' efforts, and thus the probability of spoken errors, all the words that the speaker needs to say, like "dash" "slash" and "spacebar" are included in the script.

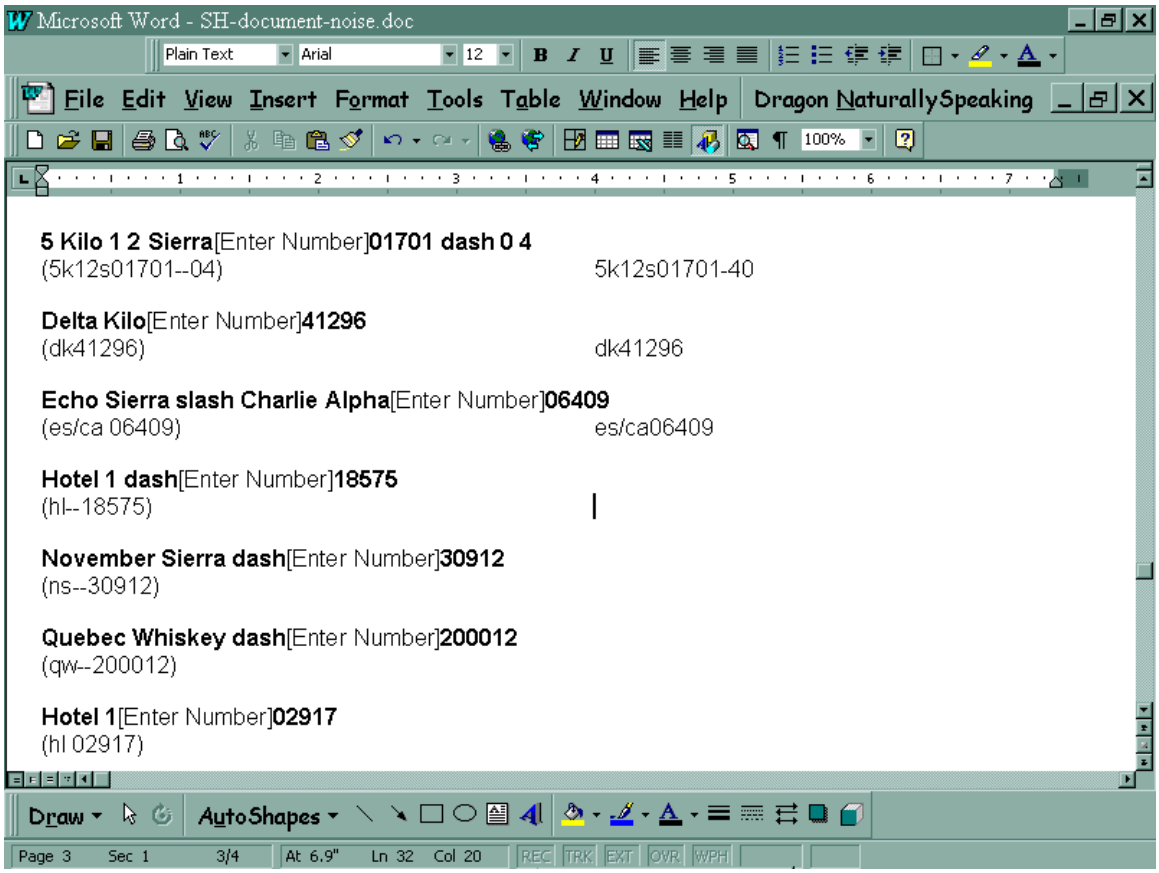

Figure 3. What the Data Entry tester sees when testing NIH Grant Numbers.

To improve accuracy, the active vocabularies of the two systems were limited, to the extent possible, to only those words and macros needed by the test scenario. For the Discrete Speech Recognition program, a special vocabulary group was created containing only the terms needed for the testing. For the Continuous Speech Recognition program, a special topic was created. Terms needed for the testing were added to the topic and as many other words as possible removed, leaving 1652 words. The Continuous Speech Recognition program does not have a mechanism for creating a very small vocabulary.

Following the tests, the output text file created via speech was compared to an error-free template using a difference program. Errors were counted and the elapsed time calculated from the timestamps.

3. RESULTS

Two measures of performance are used to evaluate Discrete and Continuous Speech Recognition technologies: accuracy and speed.

For Program Control, accuracy is defined as the number of correctly spoken and recognized commands divided by the total number of commands tested (31), expressed in percent. Unlike some evaluations of speech recognition systems [14], all error types, both misspoken and misrecognized, are weighted equally. The reason is that 1) we are evaluating the accuracy of the system in its application environment (a simulation thereof), and 2) the cost of recovery from any error type would be approximately the same.

Speed is defined as the total number of correctly recognized commands divided by the total amount of time taken to say all of the commands in the test, expressed as number of commands per minute. Figures 4 and 5 show results from the Program Control tests. The sample size for each case is 20.

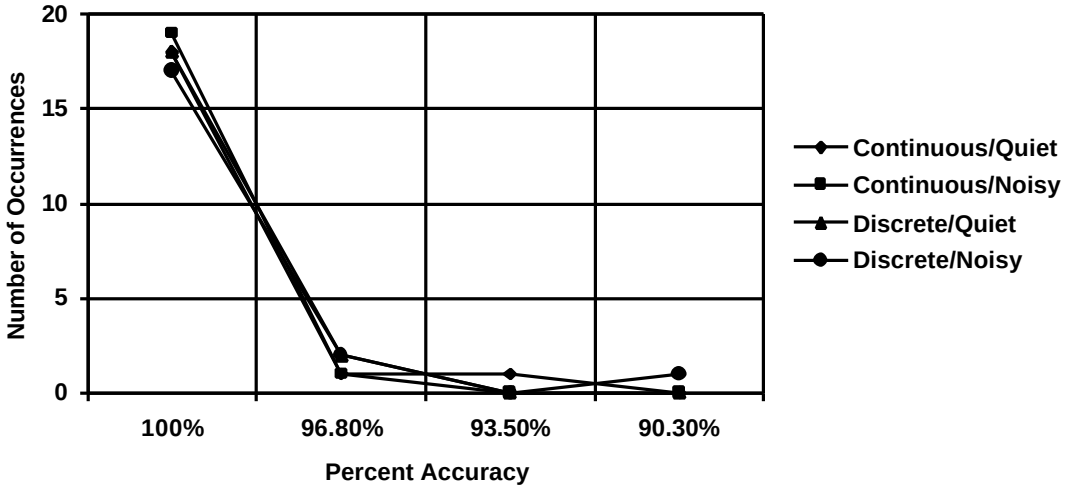

Figure 4. Program Control accuracy.

In all scenarios, accuracy was 100% for most of the participants. T tests (confidence interval of 90%, $P > 0.05$) indicate there is no significant difference between Quiet and Noisy results for either the Continuous Speech Recognition or Discrete Speech Recognition program, and no significant difference between the Continuous and Discrete Speech Recognition program results for either Quiet or Noisy environments.

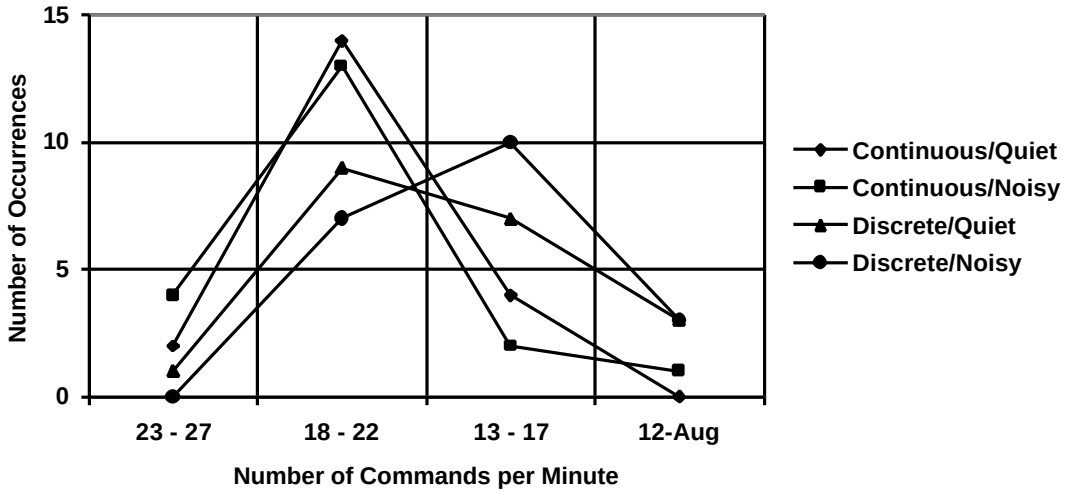

Figure 5. Program Control speed.

Speed data show more variability. T tests (confidence interval of 90%, $P > 0.05$) indicate there is no significant difference between Quiet and Noisy results for either the Continuous or Discrete Speech Recognition program. T tests do indicate that the measured differences in speed between the Continuous

Speech Recognition and Discrete Speech Recognition programs for both Quiet and Noisy environments are significant, with the Continuous Speech Recognition program being about 14% faster. Both technologies are sufficiently fast to control the scan application.

For Data Entry, accuracy is calculated as the number of correctly recognized strings of words, numbers and letters divided by the total number of strings tested (40), expressed in percent. Because the application we are targeting requires an complete string to be entered correctly, we are interested in measuring its ability to recognize the entire string, rather than individual utterances[14]. Thus, if any part of the string was recognized incorrectly, the entire string was considered incorrect. Speed is calculated as the number of correctly recognized complete strings divided by the total amount of time taken to say all of the strings in the test, expressed as number of strings per minute. Figures 6 and 7 show results from the Data Entry tests. The sample size for each case is 20.

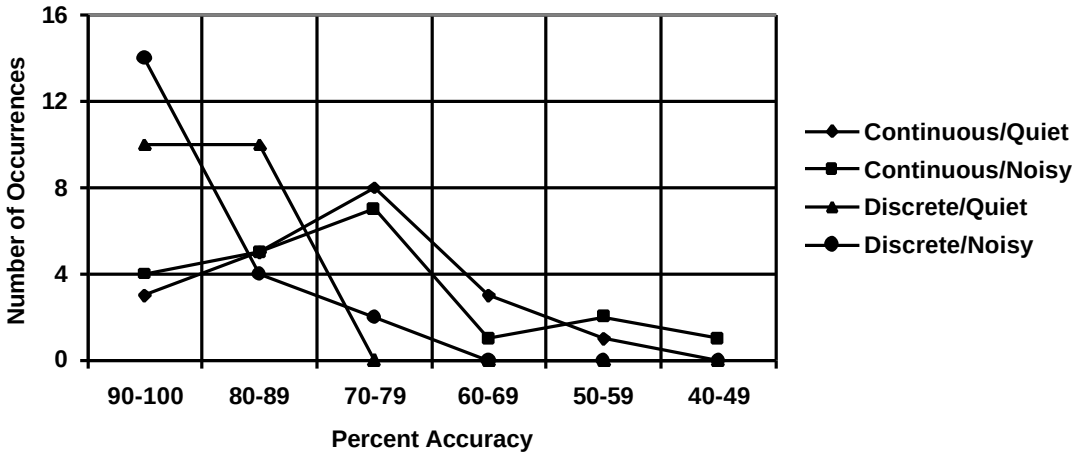

Figure 6. Data Entry accuracy.

The apparent difference between the accuracy of the Continuous and Discrete Speech Recognition programs for Data Entry is confirmed by T tests (confidence interval of 90%, $P > 0.05$). There is a significant difference in accuracy between the Continuous and Discrete Speech Recognition programs in both Quiet and Noisy environments, with the Discrete Speech Recognition program being about 90% accurate and the Continuous Speech Recognition program being about 77% accurate. There is no significant difference between Quiet and Noisy results for either the Continuous Speech Recognition or Discrete Speech Recognition program. We observed that the Continuous Speech Recognition program was prone to insert extra characters into the alphanumeric strings. Because it expects continuous utterances, we hypothesize that the Continuous Speech Recognition program recognizes breaths and other non-speech sounds as additional characters. Because the Discrete Speech Recognition program requires separated speech, it does not attempt to interpret small extraneous sounds, resulting in greater accuracy for this task.

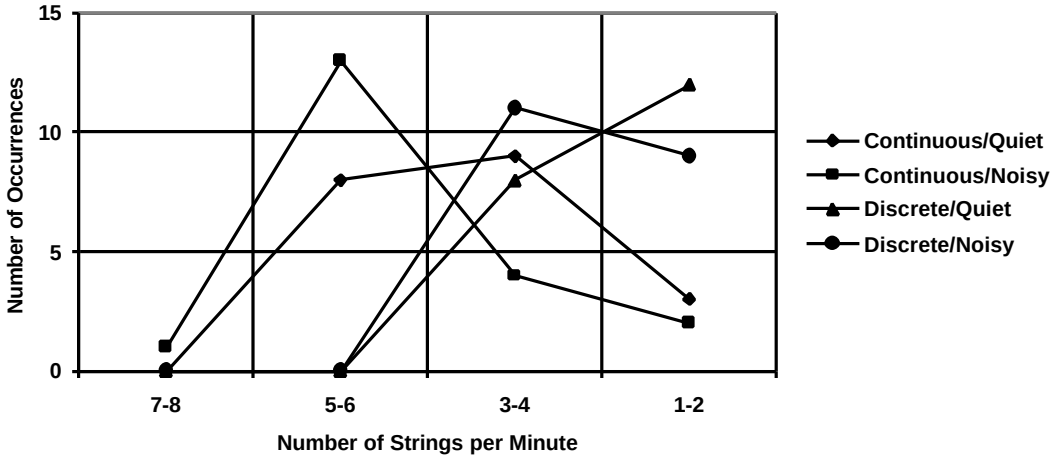

Figure 7. Data Entry speed.

The apparent difference in speed between the Continuous and Discrete Speech Recognition programs is confirmed by T tests (confidence interval of 90%, $P > 0.05$) as being statistically significant. The Discrete Speech Recognition program is inherently slower for Data Entry because each word (including individual numbers or alpha-bravo letters) must be separated by a short pause. Although a string of numbers can be spoken without pause, it must be preceded by the phrase "enter number." If the speaker should pause while speaking a string of numbers, the "enter number" modality ends. Some speakers demonstrated difficulty with this switch between discrete and continuous while speaking one string of words, letters and numbers. So even though Discrete Speech Recognition was more accurate, it was slower than Continuous Speech Recognition for the Data Entry tests.

The Mars-1 data entry operators type about 3 NIH Grant Numbers or Databank Accession Numbers per minute, including time to repair mistakes. The Discrete Speech Recognition program speed is comparable, but does not include the time needed to correct errors for about 10% of the strings. The Continuous Speech Recognition program is faster, but almost one fourth of the strings would have to be corrected. Both Speech Recognition speeds were optimized in the tests by prompting the testers with the correct military vocabulary word to speak for each letter entered. When combined with the time to correct mistakes, plus the added time to remember the military vocabulary, Speech Recognition is not as fast as typing for the entry of NIH Grant Numbers and Databank Accession Numbers. Learning the military vocabulary would also add to the initial training time.

The results of the tests indicate that for computer control both Continuous and Discrete Speech Recognition technology provide high accuracy and sufficient speed when the active vocabulary is limited to a small number of choices. Both perform equally well for computer control under conditions of high ambient noise.

For entering strings of words, letters and numbers such as NIH Grant Numbers or Databank Accession Numbers, neither Discrete nor Continuous Speech Recognition technology currently provide sufficient speed and accuracy to compete with a competent typist. Both were reasonably fast, but generated too many errors. Although some of the errors in the tests were due to incorrect or inarticulate speaking, these are just as likely to happen in a production environment, requiring additional time to correct.

4. IMPLEMENTATION AND FUTURE WORK

Discrete Speech Recognition has been installed on two scan workstations in the MARS-2 test system in the CEB lab. A vocabulary has been associated with the scan application that includes commands corresponding to the controls available in the scan application. Examples of these commands are "scan", "next page", and "insert a blank page". Each command triggers a macro that executes the corresponding

action. In this environment spoken commands occasionally take many seconds to be recognized, or are not recgnized at all. Since this was not a problem during the original tests, it may be the result of sharing computing resources with another resource-intensive program. As a first step in addressing the reliability issue, we will install the Continuous Speech Recognition product and determine if that improves the reliability. A follow-on strategy is to explore imbedding speech recognition directly into the scan application. Dragon Systems and other speech recognition vendors offer toolkits to support the integration of speech recognition technology into C/C++ applications. Integrating speech recognition would permit, for example, audible feedback (text-to-speech, or just beeps) to alert the operator when an utterance is not understood, or resetting the speech recognition engine if too much time passes between an utterance and recognition.

Even though Speech Recognition is fast and accurate in the presence of noise, it remains to be seen in a production environment as to whether it will be a successful adjunct to the MARS-2 scan program. Earlier we had introduced such an adjunct in the MARS-1 system, namely a foot pedal switch as an option to the mouse for initializing a scan. But after several weeks trial, the operators chose not to use it, finding the mouse to be more convenient. Similarly, operators may find speech recognition less convenient, at least initially, for the following reasons: the headset may be uncomfortable; the headset cord may be in the way; because the hands no longer need to move away from the journal, recognition may be perceived as being slow; the speech recognition engine may occasionally hang for several seconds; speaking to a computer may be perceived as "unnatural". All these reasons will have to be taken seriously and will be addressed if the operators resist adopting the new technology.

5. SUMMARY

We have concluded that for computer control both Continuous and Discrete Speech Recognition technology provide high accuracy and sufficient speed when the active vocabulary is limited to a small number of choices. Both performed well in our test environment, even under conditions of high ambient noise. Discrete Speech Recognition was installed it in the scan workstations in the MARS-2 test system in our lab. Running speech recognition concurrently with the scan program has raised some reliability issues that will be solved before speech recognition is introduced to the production environment. Our tests indicated that neither Discrete nor Continuous Speech Recognition was sufficiently fast or accurate to be used for entering strings of words, letters and numbers.

REFERENCES

1. Thoma GR and Le DX, "Medical database input using integrated OCR and document analysis and labeling technology." ."*Proceedings of the 1997 Symposium on Document Image Understanding Technology*, College Park, MD: University of Maryland Institute for Advances in Computer Studies, pages 180-181, 1997.

2. Dragon Systems, Inc. web site: http://www.dragonsys.com/.

3. Lernout and Hauspie Speech Products web site: http://www.lhs.com/

4. IBM ViaVoice web site: http://www.software.ibm.com/speech/

5. Microsoft speech recognition research web site: http://www.research.microsoft.com/srg/sproject.htm

6. Karl LR, Petty M, Shneiderman B. Speech versus mouse commands for word processing: an empirical evaluation. *International Journal of Man-Machine Studies*, Vol. 39, 1993, pp. 667-687.

7. Personal voice-users web sites: http://www.out-loud.com and http://idt.net/~edrose19/page7.html.

8. Zemmel NJ, Park SM, Maurer EJ, Leslie LF, Edlich RF. Evaluation of voicetype dictation for Windows for the radiologist. *Medical Progress through Technology*, Vol. 21, 1997, pp. 177-180.

9. Swett HA, Mutalik PG, Neklesa VP, Horvath L, Lee C, Richter J, Tocino I, Fisher PR. Voice-activated retrieval of mammography reference images. *Journal of Digital Imaging*, Vol. 11, 1998, pp. 65-73.

10. Nakatsu R, Susuki Y. What does voice-processing technology support today? *Proceedings of the National Academy of Sciences of the USA*, Vol. 92, 1995, pp. 10023-10030.

11. Schwartz LH, Kijewski P, Hertogen H., Roossin PS, Castellino RA. Voice recognition in radiology reporting. *American Journal of Roentgenology*, Vol. 169, 1997, pp. 27-29.

12. Voice-users Maillist archives: http://voicerecognition.com/voice-users/archive/1998/

13. Brereton P. Noise Level Evaluation of Area, Bldg. 38A/B1-N30B, for Planned Installation of Speech Recognition Computers. Internal report, National Library of Medicine, Bethesda, MD, 1998.

14. Young SJ, Chase LL. Speech recognition evaluation: a review of the U.S.CSR and LVCSR programmes. *Computer Speech and Language*, Vol. 12, 1998, pp. 263-279.